CHAPTER 1

Big Data

"Big data is a term used to describe data that has massive volume, comes in a variety of structures, and is generated at high velocity. This kind of data poses challenges to the traditional RDBMS systems used for storing and processing data. Bid data is paving way for newer approaches of processing and storing data."

In this chapter, we will talk about big data basics, sources, and challenges. We will introduce you to the three Vs (volume, velocity, and variety) of big data and the limitations that traditional technologies face when it comes to handling big data.

Getting Started

Big data, along with *cloud*, *social*, *analytics*, and *mobility*, are buzz words today in the information technology world. The availability of the Internet and electronic devices for the masses is increasing every day. Specifically, smartphones, social networking sites, and other data-generating devices such as tablets and sensors are creating an explosion of data. Data is generated from various sources in various formats such as video, text, speech, log files, and images. A single second of a high-definition (HD) video generates 2,000 times more bytes than that of a single page of text.

Consider the following statistics about Facebook, as reported on the company's web site:

1. There were 968 million daily active users on average for June of 2015. There were 844 million mobile daily active users on average for June of 2015.
2. There were 1.49 billion monthly active users as of June 30, 2015. There were 1.31 billion mobile monthly active users as of June 30, 2015.
3. There were 4.5 billion likes generated daily as of May 2013, which is a 67 percent increase from August 2012.

Figure 1-1 depicts the statistics of Twitter.

© Shakuntala Gupta Edward, Navin Sabharwal 2015
S.G. Edward and N. Sabharwal, *Practical MongoDB*, DOI 10.1007/978-1-4842-0647-8_1

www.it-ebooks.info

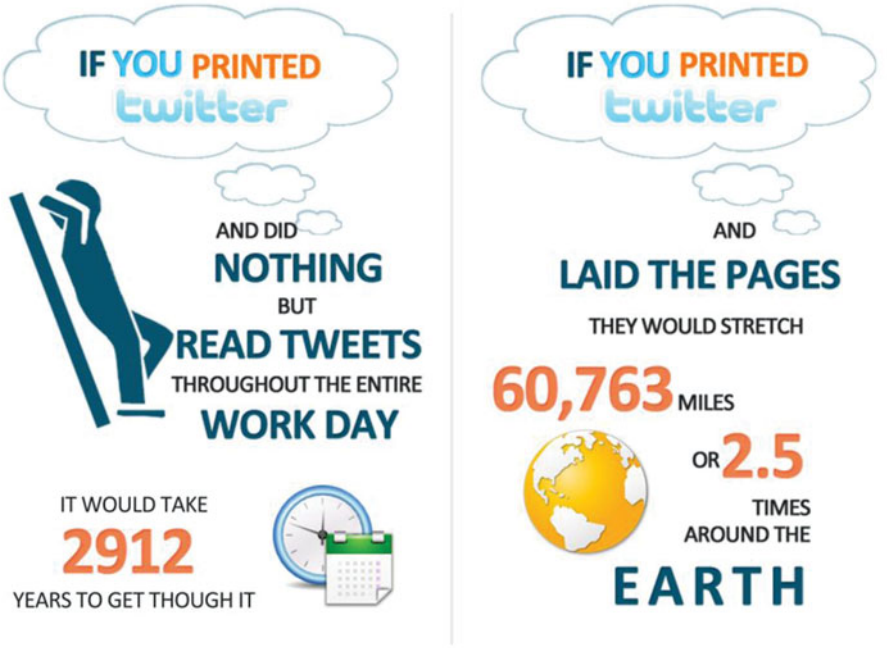

Figure 1-1. *If you printed Twitter...*

Here's another example: consider the amount of data that a simple event like going to a movie can generate. You start by searching for a movie on movie review sites, reading reviews about that movie, and posting queries. You may tweet about the movie or post photographs of going to the movie on Facebook. While travelling to the theater, your GPS system tracks your course and generates data.

You get the picture: smartphones, social networking sites, and other media are creating flood of data for companies to process and store. When the size of data poses challenges to the ability of typical software tools to capture, process, store, and manage data, then we have big data in hand. Figure 1-2 graphically defines big data.

www.it-ebooks.info

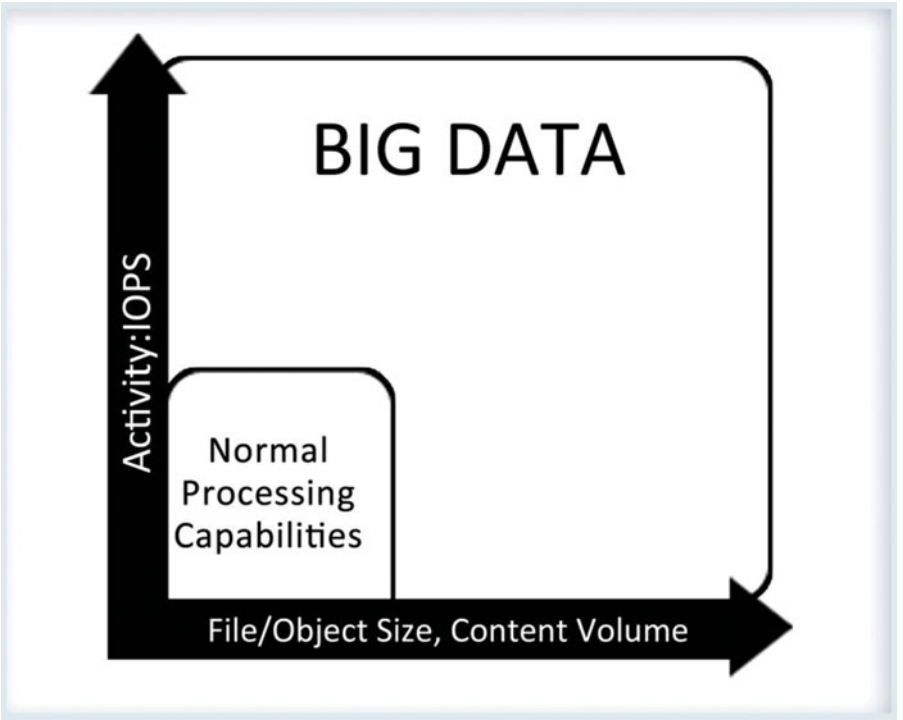

Figure 1-2. *Definition of Big Data*

Big Data

Big data is data that has high volume, is generated at high velocity, and has multiple varieties. Let's look at few facts and figures of big data.

Facts About Big Data

Various research teams around the world have done analysis on the amount of data being generated. For example, IDC's analysis revealed that the amount of digital data generated in a single year (2007) is larger than the world's total capacity to store it, which means there is no way in which to store all of the data that is being generated. Also, the rate at which data is getting generated will soon outgrow the rate at which data storage capacity is expanding.

The following sections cover insights from the MGI (McKinsey Global Institute) report (`www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)that` was published in May 2011. The study makes the case that the business and economic possibilities of big data and its wider implications are important issues that business leaders and policy makers must tackle.

The Size of Big Data Varies Across Sectors

The growth of big data is a phenomenon that is observed in every sector.MGI estimates that enterprises around the world used more than 7 exabytes of incremental disk drive data storage capacity in 2010; what's interesting is that nearly 80 percent of that total seemed to duplicate data that was stored elsewhere.

www.it-ebooks.info

MGI also estimated that, by 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per company and that many sectors had more than 1 petabyte in mean stored data per company.

Some sectors exhibited far higher levels of data intensity than others; in this case, data intensity refers to the average amount of data getting accumulated across companies/firms of that sector, implying that they have more potential to capture value from big data.

Financial services sectors, including banking, investment, and securities services, are highly transaction-oriented; they are also required by regulations to store data. The analysis shows that they have the most digital data stored per firm on average.

Communications and media firms, utilities, and government also have significant digital data stored per enterprise or organization, which appears to reflect the fact that such entities have a high volume of operations and multimedia data.

Discrete and process manufacturing have the highest aggregate data stored in bytes. However, these sectors rank much lower in intensity terms, since they are fragmented into a large number of firms.

The Big Data Type Varies Across Sectors

The MGI research also shows that the type of data stored also varies by sector. For instance, retail and wholesale, administrative parts of government, and financial services all generate significant amounts of text and numerical data including customer data, transaction information, and mathematical modeling and simulations. Sectors such as manufacturing, health care, media and communications are responsible for higher percentages of multimedia data. And image data in the form of X-rays, CT, and other scans dominate data storage volumes in health care.

In terms of geographic spread of big data, North America and Europe have 70% of the global total currently. Thanks to cloud computing, data generated in one region can be stored in another country's datacenter. As a result, countries with significant cloud and hosting provider offerings tend to have high storage of data.

Big Data Sources

In this section, we will cover the major factors that are contributing to the ever increasing size of data. Figure 1-3 depicts the major contributing sources.

www.it-ebooks.info

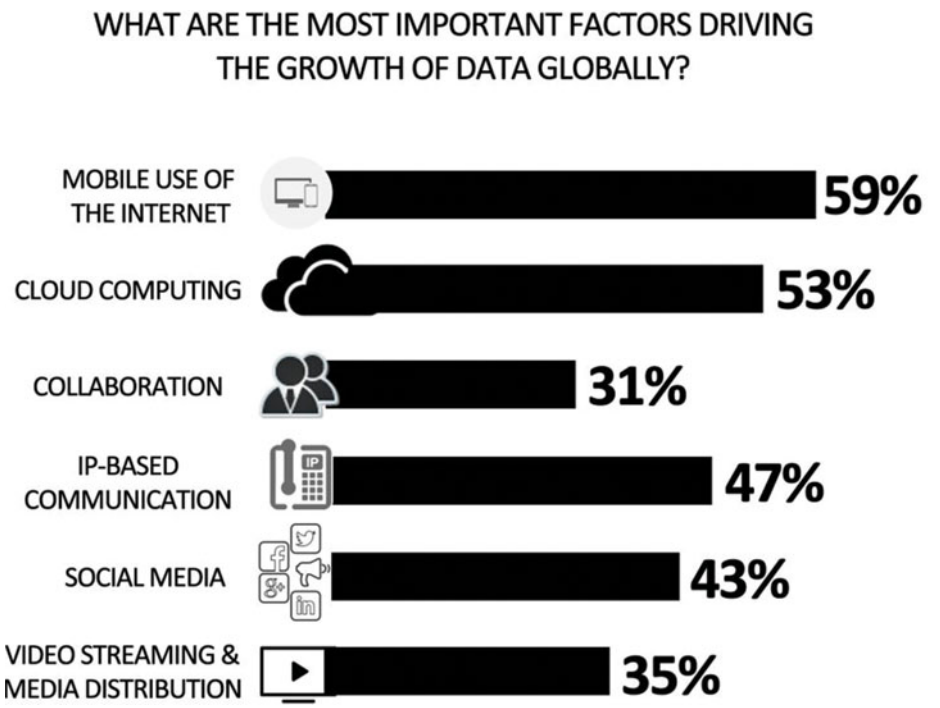

Figure 1-3. *Sources of data*

As highlighted in the MGI report, the major sources of this data are

- Enterprises, which are collecting data with more granularities now, attaching more details with every transaction in order to understand consumer behavior.
- Increase in multimedia usage across industries such as health care, product companies, etc.
- Increased popularity of social media sites such as Facebook, Twitter, etc.
- Rapid adoption of smartphones, which enable users to actively use social media sites and other Internet applications.
- Increased usage of sensors and devices in the day-to-day world, which are connected by networks to computing resources.

The MGI report also projects that the number of machine-to-machine devices such as sensors (which are also referred as the Internet of Things, or IoT) will grow at a rate exceeding 30 percent annually over the next five years.

Thus, the rate of growth of data is increasing and so is the diversity. Also, the model of data generation has changed from few companies generating data and others consuming it to everyone generating data and everyone consuming it. This is due to the penetration of consumer IT and internet technologies along with trends like social media. Figure 1-4 depicts the change in the data generation model.

www.it-ebooks.info

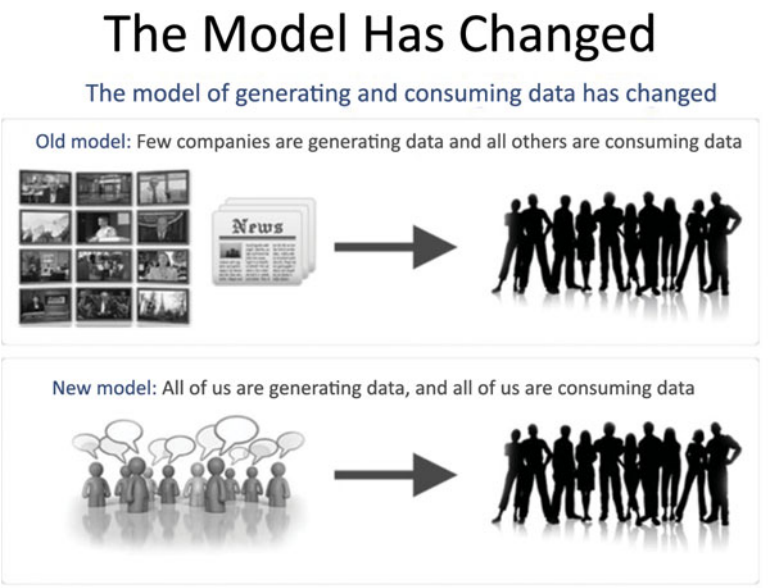

Figure 1-4. *Data model*

Three Vs of Big Data

We have defined big data as data with three Vs:volume, velocity, and variety, as shown in Figure 1-5. Let's look at the three Vs. It is imperative that organizations and IT leaders focus on these aspects.

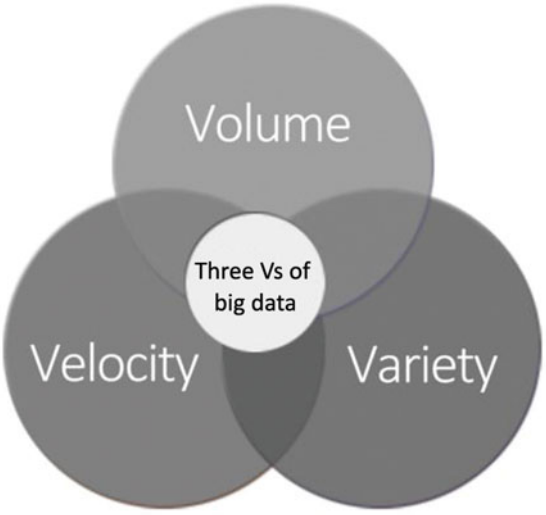

Figure 1-5. *The three Vs of big data. The "big" isn't just the volume*

www.it-ebooks.info

Volume

Volume in big data means the size of the data. As discussed in the previous sections, various factors contribute to the size of big data: as businesses are becoming more transaction-oriented, we see ever increasing numbers of transactions; more devices are getting connected to the Internet, which is adding to the volume; there is an increased usage of the Internet; and there is an increase in the digitization of content. Figure 1-6 depicts the growth in digital universe since 2009.

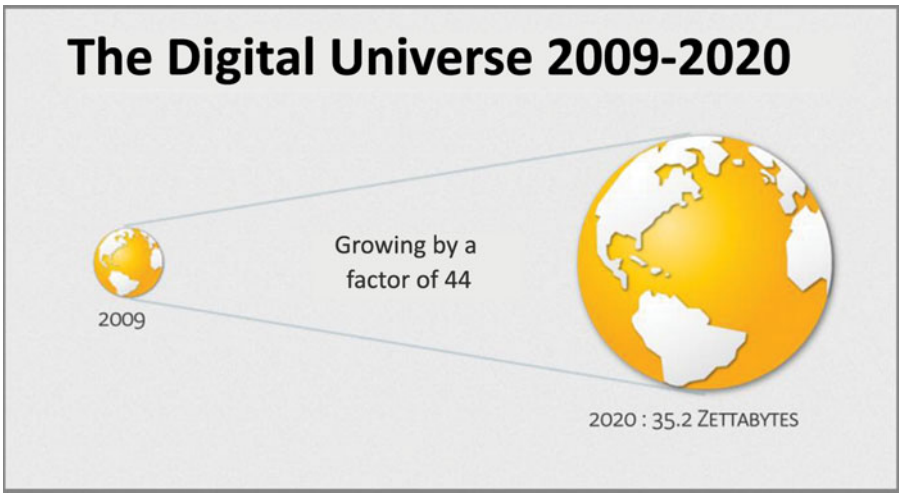

Figure 1-6. *Digital universe size*

In today's scenario, data is not just generated from within the enterprise; it's also generated based on transactions with the extended enterprise and customers. This requires extensive maintenance of customer data by the enterprises. A petabyte scale is becoming commonplace these days. Figure 1-7 depicts the data growth rate.

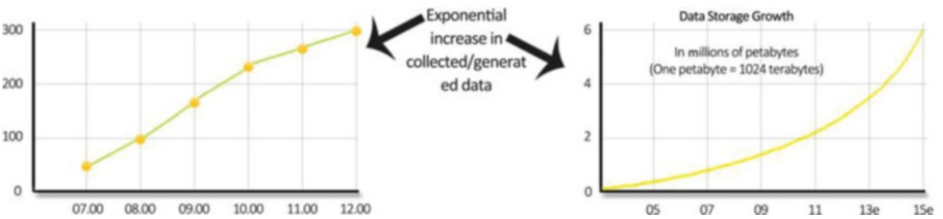

Figure 1-7. *Growth rate*

This huge volume of data is the biggest challenge for big data technologies. The storage and processing power needed to store, process, and make accessible the data in a timely and cost effective manner is massive.

www.it-ebooks.info

Variety

The data generated from various devices and sources follows no fixed format or structure. Compared to text, CSV or RDBMS data varies from text files, log files, streaming videos, photos, meter readings, stock ticker data, PDFs, audio, and various other unstructured formats.

There is no control over the structure of the data these days. New sources and structures of data are being created at a rapid pace. So the onus is on technology to find a solution to analyze and visualize the huge variety of data that is out there. As an example, to provide alternate routes for commuters, a traffic analysis application needs data feeds from millions of smartphones and sensors to provide accurate analytics on traffic conditions and alternate routes.

Velocity

Velocity in big data is the speed at which data is created and the speed at which it is required to be processed. If data cannot be processed at the required speed, it loses its significance. Due to data streaming in from social media sites, sensors, tickers, metering, and monitoring, it is important for the organizations to speedily process data both when it is on move and when it is static (see Figure 1-8). Reacting and processing quickly enough to deal with the velocity of data is one more challenge for big data technology.

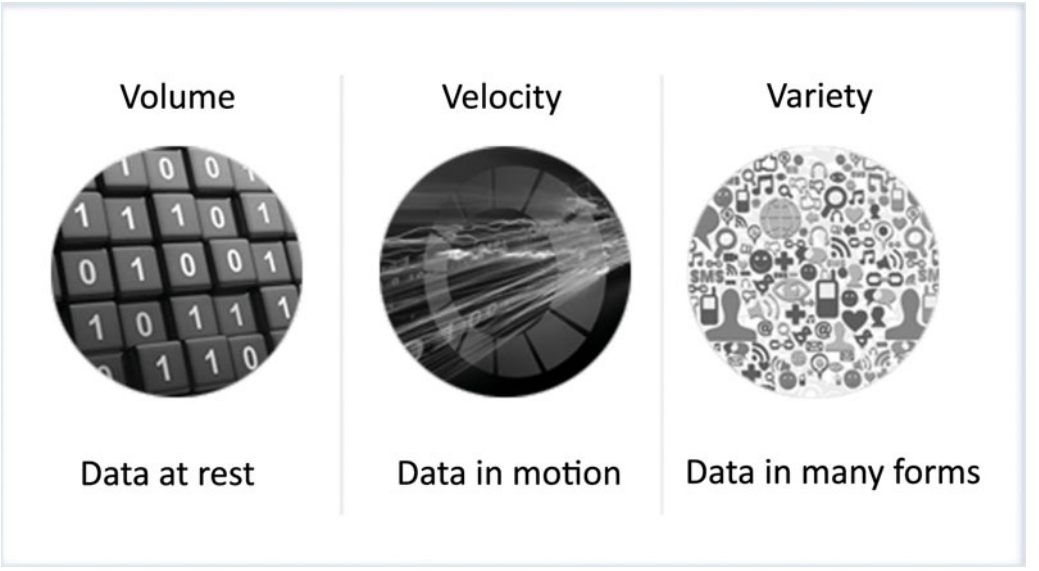

Figure 1-8. *The three aspects of data*

Real-time insight is essential in many big data use cases. For example, an algorithmic trading system takes real-time feeds from the market and social media sites like Twitter to make decisions on stock trading. Any delay in processing this data can mean millions of dollars in lost opportunities on a stock trade.

There is a fourth V that is talked about whenever big data is discussed. The fourth V is veracity, which means not all the data out there is important, so it's essential to identify what will provide meaningful insight, and what should be ignored.

www.it-ebooks.info

Usage of Big Data

This section will focus on ways of using big data for creating value for organizations. Before we delve into how big data can be made usable to the organizations, let's first look at why big data is important.

Big data is a completely new source of data; it's data that is generated when you post on a blog, like a product, or travel. Previously, such minutely available information was not captured. Now it is and organizations that embrace such data can pursue innovations, improve their agility, and increase their profitability.

Big data can create value for any organization in a variety of ways. As listed in the MGI report, this can be broadly categorized into five ways of usage of big data.

Visibility

Accessibility to data in a timely fashion to relevant stakeholders generates a tremendous amount of value. Let's understand this with an example. Consider a manufacturing company that has R&D, engineering, and manufacturing departments dispersed geographically. If the data is accessible across all these departments and can be readily integrated, it can not only reduce the search and processing time but will also help in improving the product quality according to the present needs.

Discover and Analyze Information

Most of the value of big data comes from when the data collected from outside sources can be merged with the organization's internal data. Organizations are capturing detailed data on inventories, employees, and customers. Using all of this data, they can discover and analyze new information and patterns; as a result, this information and knowledge can be used to improve processes and performance.

Segmentation and Customizations

Big data enables organizations to create tailor-made products and services to meet specific segment needs. This can also be used in the social sector to accurately segment populations and target benefit schemes for specific needs. Segmentation of customers based on various parameters can aid in targeted marketing campaigns and tailoring of products to suit the needs of customers.

Aiding Decision Making

Big data can substantially minimize risks, improve decision making, and uncover valuable insights. Automated fraud alert systems in credit card processing and automatic fine-tuning of inventory are examples of systems that aid or automate decision-making based on big data analytics.

Innovation

Big data enables innovation of new ideas in the form of products and services. It enables innovation in the existing ones in order to reach out to large segments of people. Using data gathered for actual products, the manufacturers can not only innovate to create the next generation product but they can also innovate sales offerings.

As an example, real-time data from machines and vehicles can be analyzed to provide insight into maintenance schedules; wear and tear on machines can be monitored to make more resilient machines; fuel consumption monitoring can lead to higher efficiency engines. Real-time traffic information is already making life easier for commuters by providing them options to take alternate routes.

www.it-ebooks.info

Thus, big data is not just the volume of data. It's the opportunities in finding meaningful insights from the ever-increasing pool of data. It's helping organizations make more informed decisions, which makes them more agile. It not only provides the opportunity for organizations to strengthen existing business by making informed decisions, it also helps in identifying new opportunities.

Big Data Challenges

Big data also poses some challenges. In this section, we will highlight a few of them.

Policies and Procedures

As more and more data is gathered, digitized, and moved around the globe, the policy and compliance issues become increasingly important. Data privacy, security, intellectual property, and protection are of immense importance to organizations.

Compliance with various statutory and legal requirements poses a challenge in data handling. Issues around ownership and liabilities around data are important legal aspects that need to be dealt with in cases of big data.

Moreover, many big data projects leverage the scalability features of public cloud computing providers. This poses a challenge for compliance.

Policy questions on who owns the data, what is defined as fair use of data, and who is responsible for accuracy and confidentiality of data also need to be answered.

Access to Data

Accessing data for consumption is a challenge for big data projects. Some of the data may be available to third parties, and gaining access can be a legal, contractual challenge.

Data about a product or service is available on Facebook, Twitter feeds, reviews, and blogs, so how does the product owner access this data from various sources owned by various providers?

Likewise, contractual clauses and economic incentives for accessing big data need to be tied in to enable the availability of data by the consumer.

Technology and Techniques

New tools and technologies built specifically to address the needs of big data must be leveraged, rather than trying to address the aforementioned issues through legacy systems. The inadequacy of legacy systems to deal with big data on one hand and the lack of experienced resources in newer technologies is a challenge that any big data project has to manage.

Legacy Systems and Big Data

In this section, we will discuss the challenges that organizations are facing when managing big data using legacy systems.

Structure of Big Data

Legacy systems are designed to work with structured data where tables with columns are defined. The format of the data held in the columns is also known.

www.it-ebooks.info

However, big data is data with many structures. It's basically unstructured data such as images, videos, logs, etc.

Since big data can be unstructured, legacy systems created to perform fast queries and analysis through techniques like indexing based on particular data types held in various columns cannot be used to hold or process big data.

Data Storage

Legacy systems use big servers and NAS and SAN systems to store the data. As the data increases, the server size and the backend storage size has to be increased. Traditional legacy systems typically work in a scale-up model where more and more compute, memory, and storage needs to be added to a server to meet the increased data needs. Hence the processing time increases exponentially, which defeats the other important requirement of big data, which is velocity.

Data Processing

The algorithms in legacy system are designed to work with structured data such as strings and integers. They are also limited by the size of data. Thus, legacy systems are not capable of handling the processing of unstructured data, huge volumes of such data, and the speed at which the processing needs to be performed.

As a result, to capture value from big data, we need to deploy newer technologies in the field of storing, computing, and retrieving, and we need new techniques for analyzing the data.

Big Data Technologies

You have seen what big data is. In this section we will briefly look at what technologies can handle this humongous source of data. The technologies in discussion need to efficiently accept and process different types of data.

The recent technology advancements that enable organizations to make the most of its big data are the following:

1. New storage and processing technologies designed specifically for large unstructured data
2. Parallel processing
3. Clustering
4. Large grid environments
5. High connectivity and high throughput
6. Cloud computing and scale-out architectures

There are a growing number of technologies that are making use of these technological advancements. In this book, we will be discussing MongoDB, one of the technologies that can be used to store and process big data.

www.it-ebooks.info

Summary

In this chapter you learned about big data. You looked into the various sources that are generating big data, and the usage and challenges posed by big data. You also looked why newer technologies are needed to store and process big data.

In the following chapters, you will look into a few of the technologies that help organizations manage big data and enable them to get meaningful insights from big data.

www.it-ebooks.info